Developmental Science

Developmental Science 16:2 (2013), pp 149–158 DOI: 10.1111/desc.12011

PAPER

Probabilistic cue combination: less is more

Daniel Yurovsky,[1] Ty W. Boyer,[2] Linda B. Smith[3] and Chen Yu[3]

1. *Department of Psychology, Stanford University, USA*
2. *Department of Psychology, Georgia Southern University, USA*
3. *Department of Psychological and Brain Sciences and Program in Cognitive Science, Indiana University, USA*

Abstract

Learning about the structure of the world requires learning probabilistic relationships: rules in which cues do not predict outcomes with certainty. However, in some cases, the ability to track probabilistic relationships is a handicap, leading adults to perform non-normatively in prediction tasks. For example, in the dilution effect*, predictions made from the combination of two cues of different strengths are less accurate than those made from the stronger cue alone. Here we show that* dilution *is an adult problem; 11-month-old infants combine strong and weak predictors normatively. These results extend and add support for the* less is more *hypothesis: limited cognitive resources can lead children to represent probabilistic information differently from adults, and this difference in representation can have important downstream consequences for prediction.*

Introduction

Succeeding in the world requires making accurate predictions. From patterns of light on the retina, one must predict the structure of the environment. From a set of job applicants, one must predict which will be the best employee. From a set of potential foods, one must predict which will result in a delicious dinner and which will result in an upset stomach. These are difficult problems because they involve probabilistic relationships: no cue predicts the desired outcome with 100% certainty. Probabilistic prediction is a general problem faced by cognitive systems (Brunswik, 1943; Ramscar, Yartlett, Dye, Denny & Thorpe, 2010).

Humans are remarkably good at this. For instance, Griffiths and Tenenbaum (2006) showed that the average undergraduate can predict movie runtimes, lengths of poems, and reigns of pharaohs with high accuracy from a single piece of information. Even young infants are able to track (Saffran, Aslin & Newport, 1996) and make predictions (Xu & Garcia, 2008) from probabilistic events in their environments. These processes appear across tasks (e.g. visual perception: Kersten & Yuille, 2003; motor control: Körding & Wolpert, 2006; memory retrieval: Shiffrin & Steyvers, 1997) and are available quite early (e.g. newborns: Bulf, Johnson & Valenza, 2011; 2-month-olds: Kirkham, Slemmer & Johnson, 2002, 6-month-olds: Shukla, White & Aslin, 2011). Their ubiquity has inspired hope for a unified understanding of both mature and developing cognitive systems as probabilistic prediction machines (Chater, Tenenbaum & Yuille, 2006; Tenenbaum, Kemp, Griffiths & Goodman, 2011; Clark, in press). Unsurprisingly, the efficiency of these processes generally improves over development (Xu & Tenenbaum, 2007; Smith & Yu, 2008; Thiessen, 2010); however, on some probabilistic prediction tasks, young children actually outperform adults.

In perhaps the simplest such task (Derks & Paclisanu, 1967; see also Gardner, 1957), participants are presented with two lights, and, in a series of trials, must predict which light will activate. If they make the correct prediction, they receive a reward. The stimuli are probabilistic – one light activates on 70% of the trials, and the other light activates on the remaining 30%. The optimal strategy – the one that *maximizes* rewards – is to always select the more probable light. This is precisely how 3- and 4-year-old children behave. However, it is not how adults behave; adults *probability match*, selecting each light in proportion to its probability of activation, reducing their total reward (Estes, 1976). This perplexing result has been explained as a type of apophenia: search for local sequential patterns in the light sequences that

Address for correspondence: Daniel Yurovsky, Stanford University, Department of Psychology, 450 Serra Mall, Stanford, CA, 94305, USA; e-mail: yurovsky@stanford.edu

do not exist (Wolford, Newman, Miller & Wig, 2004, Yu & Cohen, 2009). Thus, adults' prowess in tracking probabilities at multiple levels leads to suboptimal performance in this task. This argument is reinforced by two further sources of evidence. First, adults who are more likely to *probability match* are also more likely to discover local pattern structure if it does exist (Gaissmaier & Schooler, 2008). Second, adults perform more normatively when they have fewer cognitive resources available; maximizing more often under dual-task conditions (Wolford *et al*., 2004; Gaissmaier, Schooler & Rieskamp, 2006).

Newport and colleagues (Johnson & Newport, 1989; Newport, 1990; Hudson Kam & Newport, 2005) have proposed that children's resource constraints in probabilistic prediction are adaptive. Under their *less is more* hypothesis, children outperform adults in learning languages precisely because their resource constraints limit their ability to entertain complex hypotheses. Elman (1993) formalized this claim, showing that initially resource-constrained neural networks learned grammatical structure better than unconstrained nets. Resource constraints prevented the search for complex patterns, keeping networks from getting stuck in local maxima. In a language-learning task analogous to the light prediction task above, Hudson Kam and Newport (2005) showed that adults *probability match* their language input, whereas 6-year-olds *maximize*, always picking the most probable alternative. Further, as before, increasing task demands lead to increased *maximizing* in adults (Hudson Kam & Newport, 2009). The *less is more* hypothesis (Newport, 1990) thus suggests that the representation of probabilistic information changes over development. If *maximizing* is a general property of young children's probability learning, then they should also outperform adults in other cases in which optimal performance results from *maximizing* rather than *probability matching*. This paper tests this prediction in the context of combining information from two probabilistic predictors of different strengths.

Across a range of domains, tasks, and developmental ages, evidence from two strong predictors leads to better learning and prediction than a single strong predictor alone (Shanteau, 1975; Ernst & Banks, 2002; McKenzie, Lee & Chen, 2002; Yoshida & Smith, 2005; Frank, Slemmer, Marcus & Johnson, 2009). However, the addition of evidence from a weaker predictor to a stronger predictor can lead to non-normative behavior in adults. For instance, in the 'bookbags-and-pokerchips' task, adults are shown two bookbags and told that one contains 70 white chips and 30 red chips, and the other contains the opposite red/white ratio. The experimenter then secretly chooses one of the bags, and randomly draws a white chip. When asked to guess the bag's identity, participants are 60% certain it is the 70 white/30 red bag. Then, a second sample is drawn – three white chips, and three red chips – and participants are again asked to guess the bag's identity. While this second sample is *nondiagnostic* (i.e. equally likely to have come from either bag), participants decrease their certainty in the white-heavy bag (Shanteau, 1975). This *dilution* effect is also found in more naturalistic settings (Nisbett, Zuckier & Lemley, 1981) and even when the additional evidence is a *weaker positive* predictor rather than a *nondiagnostic* predictor (McKenzie *et al*., 2002). But the *less is more* hypothesis predicts that young learners will combine strong and weak predictors more optimally.

The *dilution* effect should depend critically on how the strength of each predictor is represented. The addition of a *Weak Cue* can dilute evidence from a *Strong Cue* only if the cognitive system represents the strength of each cue in proportion to its probability (*probability matching*). On the other hand, if the system represents only the most probable outcome (*maximizing*), then a *Weak Cue* can only add to a *Strong Cue*. Because they are both coded as strong cues, their combination should act as an even stronger cue (e.g. Ernst & Banks, 2002; McKenzie *et al*., 2002; Frank *et al*., 2009). Thus, we predict that where adults average a strong and weak probabilistic predictor, 11-month-old infants should treat their combination additively. In the experiments that follow, two centrally presented geometric shapes differed in the probability with which they predicted the appearance of a reinforcing cartoon character stimulus. One shape was a *Strong Cue*, and the other was a *Weak Cue*. Then, after learning these shape–outcome relationships, participants were tested on trials in which *Both Cues* appeared together. We predicted that adults would *probability match* – predicting more for the *Strong Cue* than for the *Weak Cue* – and that *Both Cues* would be treated as intermediate evidence. Infants, in contrast, would *maximize* in response to both the *Strong* and *Weak Cues* – treating them similarly – and predict even more strongly in response to *Both Cues*.

Experiment 1

Method

Participants

Twenty-four undergraduate students at Indiana University, and 24 11-month-old infants (mean age 11 mo 15 days; range: 10 mo 15 days to 12 mo 7 days, 13 female) participated in the experiment. Three additional adults and 11 additional infants were excluded because of failure to calibrate, incomplete data, and (infants only) fussiness. Participants received partial course credit (adults) or a small gift (infants) as compensation.

Stimuli

Stimuli fell into two categories – *cues* and *reinforcers*. Cues were videos of monochromatic geometric shapes looming and shrinking. Four such shapes were created: a red square, a blue circle, a green triangle, and a yellow diamond. Reinforcers were videos of cartoon characters, each of which displayed a different animated behavior. For instance, one reinforcer consisted of a bouncing purple stuffed animal. Each of the three reinforcer videos was accompanied by a unique sound. Videos for all stimuli were 2 seconds long.

Design and procedure

The experiment consisted of a series of trials in which a cue appeared centrally on the screen for 2 seconds and then was followed by two blank boxes that appeared for 1 second on each side of the central location. After this, on some trials a reinforcer would appear in one of the boxes for 2 seconds. On other trials, the boxes would remain blank for 2 seconds. Figure 1 shows a schematic of an example reinforcement trial.

Each participant saw 10 such trials for each of two unique cues. One shape was a *Strong Cue* – predicting the appearance of a random reinforcer on seven of 10 trials. On the other three trials, the boxes remained blank. The other shape was a *Weak Cue* – predicting the appearance of a random reinforcer on only four of 10 trials. After 20 single-cue trials, participants saw five trials on which *Both Cues* appeared together, and which were never followed by a reinforcer.

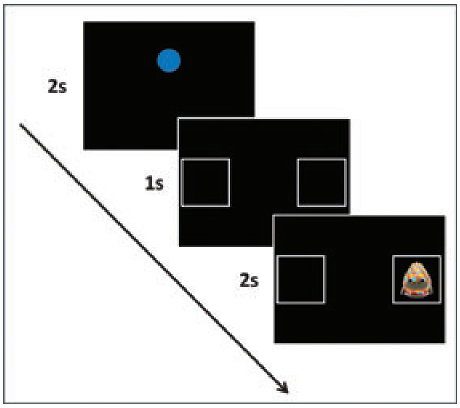

Figure 1 *A schematic of one experimental trial. A cue loomed on the screen for two seconds, was replaced by two empty boxes for 1 second, and then a reinforcer played in one of the boxes for 2 seconds.*

All reinforcers appeared on the same side of the screen for a given participant, and reinforcer sides and cue identities were counterbalanced across participants. Single-cue trials appeared in random order until all 10 trials of each cue had been seen. Finally, an attention-getter was shown prior to the onset of each trial and remained on screen until fixated for at least 100 ms.

Participants watched the experimental videos on a 17-inch monitor while their eye movements were recorded by a Tobii 1750 eye tracker (see Appendix for details). The eye tracker was calibrated for each participant before the experiment began. To facilitate fair comparison between adult and infant participants, adults were only instructed to watch the screen for the duration of the experiment.

In order to determine how cues affected participants' predictions about the appearance of reinforcers, we analyzed predictive looking after the offset of the cue, and thus, the dependent measure of interest was latency to saccade to either of the boxes. However, because reinforcers appeared in these boxes after 1 second on reinforcement trials (Figure 1), any saccades initiated after this point were more likely reactive than predictive. Allowing 200 milliseconds for saccade initiation (Engel, Anderson & Soechting, 1991), only eye movements within 1.2 seconds of cue-offset were analyzed. For a related location-based prediction paradigm, and another comparison of predictive learning in adults and infants, see Richardson and Kirkham (2004).

Results and discussion

The key empirical question is how participants' probability of predicting the appearance of the reinforcer in each box varied in the *Weak*, *Strong*, and *Both Cues* conditions. To answer this question, one wants to estimate the relative probability of looking to either the Correct or Incorrect box over time as a function of cue. To ensure that we measured predictive rather than reactive saccades, we analyzed only eye movements in the first 1.2 seconds after *cue*-offset (see above). However, not all participants made a saccade on all trials in this window, leading to *right-censored* data. Since simply excluding these trials would produce a biased estimate of saccade probability, the appropriate statistical analysis is a proportional hazards regression (Cox, 1972). This analysis estimates a *hazard function* for each condition – the probability of making a saccade at each time-point given that no saccade has yet been made (for other examples of Cox regression in developmental studies, see e.g. Zosuls, Ruble, Tamis-LeMonda, Shrout, Bornstein & Greulich, 2009; Kidd, Piantadosi & Aslin, 2012).

Cox regression is a semi-parametric model: it makes no assumptions about the functional form of the hazard function. Instead, a baseline hazard function is estimated empirically from one condition, and other conditions are assumed to have hazard functions proportional to that baseline. Here, the baseline function was estimated from the *Strong Cue* condition, both because subsequent analysis becomes most straightforward, and because it contained the greatest proportion of valid eye-tracking data. This produces the most robust function estimate. Because participants could make a saccade to one of two locations on each trial (Correct or Incorrect), the regression model was stratified by location (Lunn & McNeil, 1995). That is, cues could have different effects on the hazard rate for different locations. Models were fit separately for adults and infants using the last five trials of each single cue (*Weak* and *Strong*) condition, and all five trials of the *Both Cue* condition.

These analyses indicate that adults and infants alike were more likely to look predictively to the Correct than the Incorrect location (Adults: $\beta = -2.05$, $z = -6.72$, $p < .001$; Infants: $\beta = -.779$, $z = -3.48$, $p = .001$). However, adults and infants treated the cues differently. The adult data are shown in Figure 2a. Compared to the *Strong Cue*, both the *Weak Cue* ($\beta = -.388$, $z = -2.05$, $p < .05$) and *Both Cues* ($\beta = -.401$, $z = -2.09$, $p < .05$) elicited less predictive looking to the Correct location. The *Weak Cue* ($\beta = .77$, $z = 2.24$, $p < .05$), but not *Both Cues* ($\beta = .507$, $z = 1.41$, *ns*), produced more prediction to the Incorrect location than the *Strong Cue*. Thus, the *Strong Cue* elicited the proportionally highest correct predictions, the *Weak Cue* the lowest, and *Both Cues* were intermediate. This is evidence for a *dilution effect* in the visual domain.

The infant data (Figure 2b) show maximizing in the face of both cues and no *dilution effect*. Predictive looking to the Correct location was unaffected by cue type (*Weak vs. Strong:* $\beta = -.082$, $z = -.421$, *ns*; *Both vs. Strong:* $\beta = .091$, $z = -.475$, *ns*). But, relative to the *Strong Cue*, *Both Cues* ($\beta = -1.16$, $z = -3.19$, $p < .01$) but not the *Weak Cue* ($\beta = -.391$, $z = -.1.38$, *ns*) reduced prediction to the Incorrect (i.e. non-reinforced) location. Thus, seeing *Both Cues* significantly decreased infants' probability of making Incorrect predictions, producing a relatively higher proportion of Correct predictions. Infants *maximized* when they saw the *Weak Cue* – treating it just like the *Strong Cue* – and *Both Cues* reduced incorrect prediction.

In brief, when exposed to the same multi-modal regularities, infants and adults responded by making

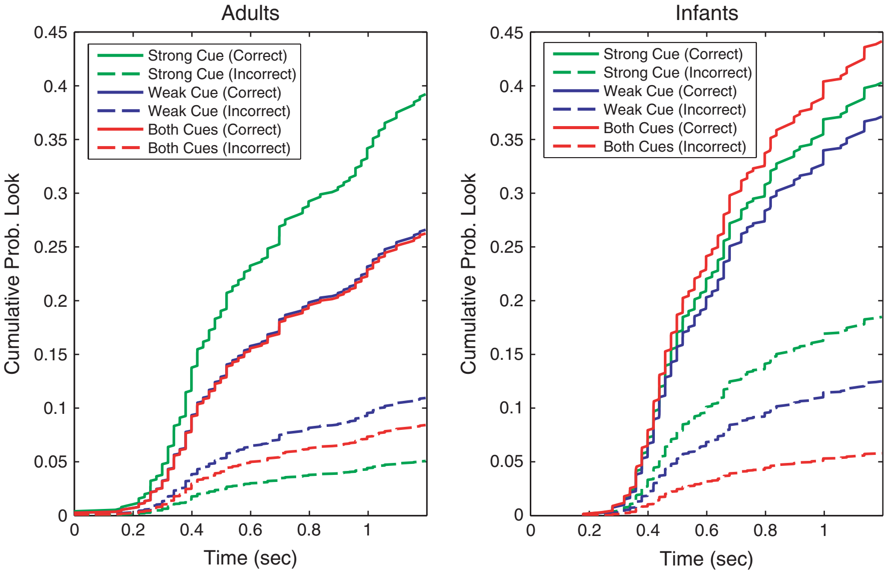

Figure 2 *Cumulative hazard functions for each cue/location combination for both groups. Each curve shows the estimated cumulative probability of a predictive look to a location (Correct/Incorrect) in the presence of a particular cue* (Strong/Weak/Both) *over time. Adults (a) were more likely to make the Correct prediction when seeing the* Strong Cue *than the* Weak Cue *or* Both Cues. *Further, the* Weak Cue *increased their probability of predicting to the Incorrect location, but* Both Cues *did not. Thus,* Both Cues *were treated as intermediate between the* Strong *and* Weak *cues, indicating dilution. Infants (b) treated the* Strong *and* Weak *cues identically, indicating that, in contrast to adults, they were maximizing. Further, they were less likely to predict to the Incorrect location when cued by* Both Cues *than the* Strong Cue. *This is evidence of additive cue combination by way of reduced prediction of incorrect alternatives.*

different predictions. Adults discriminated strongly between the Correct and Incorrect location for the *Strong Cue*, weakly between the two locations for the *Weak* cue, and showed intermediate discrimination when *Both Cues* were presented together. Infants, in contrast, predicted to the Correct side at the same rate in each condition, but predicted less often to the Incorrect location in the presence of *Both Cues*.

For both groups, the proposed interpretation of the data draws on looking to both the Correct and Incorrect locations. But why do participants ever look to the Incorrect location at all? One likely explanation is that the observed gaze behavior results not just from participants' learning in the task, but also from their expectation before coming into the experiment (or prior). On the very first trial, participants could reasonably make a prediction to either box even though they had not seen a single reinforcer. The key idea is that unobserved events should be treated not as impossible, but only as less and less likely the longer they are unencountered. Thus, each reinforcer acts not only as evidence *for* the Correct location, but also as evidence *against* the Incorrect location. When *Both Cues* are seen together, both aspects of the cues are combined. Thus, while 11-month-olds' prediction systems may be too noisy to produce faster predictions to the correct locations (as evidenced by their low ceiling-level performance in other tasks), we can see evidence of their more normative combination in the reduction of Incorrect looks.

While this account is consistent with both the adult and infant data, the infant data may have a simpler explanation. A similar pattern of looking would be observed if infants did not learn anything about the cues and the cue-specific predictive probabilities, but simply learned over the course of training that outcomes appeared in the Correct but not Incorrect locations. Because the *Both Cues* test trials occurred after 20 training trials, we would expect better prediction on these test trials than on the single cue training trials. Experiment 2 was designed to test this alternative possibility.

Experiment 2

In Experiment 2, infants were exposed to the same training trials as in Experiment 1, but training was followed by two kinds of test trials. On the first, infants were shown two *New Cues* in the same positions as the cues they had seen in training. If training led to a general preference to look to the Correct location rather than specific cue-reinforcer relationships, predictive looking on these *New Cues* trials should be similar to that observed on *Both Cues* trials in the previous experiment. In contrast, if infants learned a cue-specific predictive relationship, their looking patterns should be different, perhaps providing information about their starting point (or prior) in the absence of cue-specific information. These *New Cues* trials were subsequently followed by the original *Both Cues* trials. These trials were included to test the robustness of infants' predictive learning from the single cue trials. If infants again showed improved prediction in the face of *Both Cues* after the intervening *New Cues* trials, this would be strong evidence that they learned and combined cue-specific predictive relationships.

In order to limit fussiness and fatigue, infants were shown three *New Cues* trials and three *Both Cues* trials, resulting in a total of six as compared to the five test trials of Experiment 1.

Method

Participants

Twenty 11-month-old infants (mean age 11 mo 15 days; range: 11 mo 4 days to 12 mo 3 days, 8 female) participated in the experiment. Ten additional infants were excluded because of failure to calibrate, incomplete data, and/or fussiness. Each infant received a small gift as compensation.

Stimuli

Stimuli for Experiment 2 were identical to those used in Experiment 1.

Design and procedure

As in Experiment 1, infants were exposed to a series of trials on which a centrally located cue probabilistically predicted the appearance of a reinforcer in one of the on-screen boxes. Infants were again exposed to 20 training trials – 10 on which they saw a *Strong Cue* and 10 on which they saw a *Weak Cue*. These training trials were identical to those presented in Experiment 1. Subsequently, these training trials were followed by two kinds of test trials.

The first three test trials were *New Cues* trials on which infants saw two new cues on the screen, one in each of the locations previously occupied by the cues from training. After these trials, infants saw three *Both Cues* trials identical to those in Experiment 1. For example, if the training trials each showed either a red square or a blue circle, the *New Cues* trials would show a green triangle and a yellow diamond, and the *Both Cues* trials would show the red square and blue circle again. As in Experiment 1, reinforcer side and shape/cue type/location mappings were counterbalanced across infants.

Results and discussion

As in Experiment 1, infants' predictions were measured only during the first 1.2 seconds after *cue*-offset, and look latencies to the correct and incorrect locations for each cue type were analyzed using a proportional hazards regression. The *Strong Cue* trials were again used to estimate the baseline hazard function. Because infants received only three test trials of each type, we analyzed the last three training trials for each Cue.

Figure 3 shows infants' cumulative hazard functions for each cue type and location. As in Experiment 1, infants were more likely to look predictively to the Correct than the Incorrect location ($\beta = -.842$, $z = -2.57$, $p = .01$). As in Experiment 1, infants showed evidence of maximizing, with no effect of the *Weak Cue* on looking to either the Correct ($\beta = -.07$, $z = -.26$, *ns*) or the Incorrect location ($\beta = -.26$, $z = -.63$, *ns*). Also, as in Experiment 1, *Both Cues* did not affect infants' looking to the Correct location ($\beta = -.05$, $z = -.19$, *ns*) but significantly decreased their looking to the Incorrect location ($\beta = -1.22$, $z = -2.15$, $p < .05$). Thus, as in Experiment 1, infants maximized in the face of both single cues, and showed proportionally stronger prediction when cued by *Both Cues.*

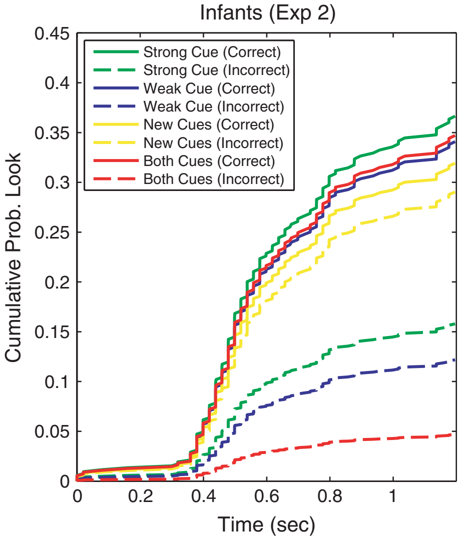

Figure 3 *Cumulative hazard functions for each cue/location combination for infants in Experiment 2. Each curve shows the estimated cumulative probability of a predictive look to a location (Correct/Incorrect) in the presence of a particular cue* (Strong/Weak/New/Both) *over time. Infants predicted the appearance of a reinforcer in the Correct location equally under all cue conditions. However, relative to the* Strong *and* Weak cues *they predicted the reinforcer on the Incorrect side less in the presence of* Both Cues *and more in the presence of the* New Cues.

When presented with the *New Cues*, infants did not alter their looking to the Correct location ($\beta = -.14$, $z = -.50$, *ns*), but were marginally *more* likely to look the Incorrect location ($\beta = .61$, $z = 1.78$, $p = .08$). As shown in Figure 3, infants showed no discrimination between the Correct and Incorrect sides in the presence of the *New Cues.* These results thus rule out the possibility that infants simply learned to look at the Correct location following the offset of any cue. Infants did not show a preference when presented with the *New Cues*, and therefore likely learned cue-specific predictive probabilities and not a general preference for the Correct side. Further, Experiment 2 shows that the result for the *Both Cues* trials is quite robust: infants showed improved prediction in the face of *Both Cues* even after the three non-reinforcing *New Cues* trials.

General discussion

Although development is generally accompanied by increased efficiency (Kail, 1991), sometimes this efficiency comes at a cost. For humans, the cost may include reduced ability to learn language (Johnson & Newport, 1989). The *less is more* hypothesis proposes that young children's resource constraints are actually critical for their success in learning language. Key evidence for this claim has taken the form of artificial language learning experiments. When exposed to inconsistent input, adults *probability match* – reproducing this inconsistency in their output. In contrast, children *maximize*, learning a simple regular pattern. The evidence presented in this paper strengthens and extends this hypothesis in two key ways. First, we show maximizing in young children in a novel domain. In addition to language learning and explicit prediction tasks (Derks & Paclisanu, 1967; Hudson Kam & Newport, 2005), maximizing is elicited even in viewing visually presented probabilistic predictors. This is strong evidence that the kind of processing critical to the *less is more* hypothesis is a general property of young learners. Second, evidence that children do not show a dilution effect suggests that the resource constraints that lead to *maximizing* have important downstream consequences. Just as in language, in which the nature of early learning can fundamentally change what is learned down the line (Elman, 1993), what is learned about probabilistic cues can fundamentally change the way that they are combined.

Following other infant prediction experiments (e.g. McMurray & Aslin, 2004; Kovács & Mehler), we designed a task to test a perceptual analogue of the *dilution effect* (Nisbett *et al*., 1981). While *dilution* is traditionally studied in explicit reasoning tasks, comparison between adults and infants necessitated construction of a perceptual paradigm. Nonetheless, our results replicate those found in explicit reasoning tasks, suggesting even more strongly that *dilution* is a fundamental property of the adult cognitive system, and licensing comparison to our younger participants (see also Knowlton, Mangels & Squire, 1996, and Gluck, Shohamy & Myers, 2002, for a similar task with adults). Whereas adult participants encoded the strengths of predictors in proportion to their probability of prediction, and subsequently combined information from them non-normatively, infants *maximized* in response to both the *Strong* and *Weak Cues* and their combination reduced prediction error. Thus, resource constraints may not only prevent children from learning probabilistic relationships that are too complex (in essence, overfitting the data – Zhu, Rogers & Gibson, 2009), they also support prediction for multiple cues that have never been experienced together (Téglás, Vul, Girotto, Gonzalez, Tenenbaum & Bonatti, 2011).

Building on other work suggesting that the statistical learning mechanisms involved in language are domain-general (Kirkham *et al*., 2002; Fiser & Aslin, 2002; Saffran, Pollack, Seibel & Shkolnik, 2007), these results suggest that more normative statistical learning in young infants may characterize other cognitive domains. For example, statistical regularities between scenes and objects play an important role in rapid object recognition (Brockmole, Castelhano & Henderson, 2006; Oliva & Torralba, 2006); cues that guide common grounding in social interactions are complex, culturally specific, and probabilistic (Bruner, 1975; Butko & Movellan, 2010; Yuki, Maddux & Masuda, 2007); noisy data about categories and category memberships often lead to rule-like over-hypotheses (Colunga & Smith, 2005; Kemp, Perfors & Tenenbaum, 2007). In all of these domains, as in language, it is interesting to ask whether the developing statistical learner might have an advantage. Could it be that for object recognition, cultural norm induction, and categorization less is also more?

Nonetheless, the adult system does develop from the infant system, and probability matching develops along with it (Derks & Paclisanu, 1967). Why develop a non-normative system? For a speculative potential explanation, we return to language. Like young children, rhesus monkeys *maximize* in response to probabilistic cues, always selecting the most likely option (Treichler, 1967; Wilson & Rolling, 1959), and when combining probabilistic predictors in a task similar to ours, monkeys perform normatively (Yang & Shadlen, 2007). That is, like infants, monkeys do not show the *dilution* effect. But, unlike young children, these monkeys will not go on to acquire language.

One of the difficulties in learning natural language is dealing with exceptions. For instance, although conjugating an English past-tense verb most often involves appending '-ed', this is not always true. In fact, some of the most frequently encountered verbs have a different character (e.g. *go* becomes *went*). Learning irregular words turns out to be quite difficult for children, who often overregularize these words (e.g. turning *go* into *goed*; Brown, 1973). Although estimates of the rates of such regularization vary, they are known to be highest before the age of 4 and to drop significantly by 7 or 8 (Marcus, Pinker, Ullman, Hollander, Rosen & Xu, 1992; Maratsos, 2000; Maslen, Theakston, Lieven & Tomasello, 2004). This seems to follow the same pattern found in the change from *probability matching* to *maximizing*. While single-mechanism accounts have been advanced that capture some of the regularities demonstrated by children in their rates and types of overregularization (e.g. Rumelhart & McClelland, 1986; Plunkett & Marchman, 1993), none successfully capture them all (MacWhinney, 1998; Pinker & Ullman, 2002). Following Elman (1993), we propose that what is missing from these accounts is developmental change. In order to deal with the regularities in language – regularities with exceptions – the cognitive system needs to represent predictors in proportion to their probabilities. Thus, it may be that for breaking into language, less is more. But, for mastering language and other complex systems, less must become more.

Acknowledgements

This research was supported by a NSF Graduate Research Fellowship to DY as well as NIH Grant R01HD056029 to CY. The authors are grateful to Celeste Kidd for help designing the stimuli, to Becca Baker, Kelsey Gibson, Amber Matthews, and Motomasa Tanioka for help with data collection, and to all of the members of the Yu, Smith, and Shiffrin labs for their feedback in discussions. We also thank Denis Mareschal and three anonymous reviewers for their feedback on this manuscript.

References

Brockmole, J.R., Castelhano, M.S., & Henderson, J.M. (2006). Contextual cueing in naturalistic scenes: global and local contexts. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **32**, 699–706.

Brown, R. (1973). *A first language*. Cambridge, MA: Harvard University Press.

Bruner, J. (1975). From communication to language: a psychological perspective. *Cognition*, **3**, 255–287.

Brunswik, E. (1943). Organismic achievement and environmental probability. *Psychological Review*, **50**, 255–272.

Bulf, H., Johnson, S.P., & Valenza, E. (2011). Visual statistical learning in the newborn infant. *Cognition*, **121**, 127–132.

Butko, N.J., & Movellan, J.R. (2010). Detecting contingencies: an infomax approach. *Neural Networks*, **23**, 973–984.

Chater, N., Tenenbaum, J.B., & Yuille, A. (2006). Probabilistic models of cognition: conceptual foundations. *Trends in Cognitive Sciences*, **10**, 287–291.

Clark, A. (in press). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences.*

Colunga, E., & Smith, L.B. (2005). From the lexicon to an expectation about kinds: a role for associative learning. *Psychological Review*, **112**, 347–382.

Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, **34**, 187–220.

Derks, P.L., & Paclisanu, M.I. (1967). Simple strategies in binary prediction by children and adults. *Journal of Experimental Psychology*, **73**, 278–285.

Elman, J.L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition*, **48**, 71–99.

Engel, K.C., Anderson, J.H., & Soechting, J.F. (1991). Oculomotor tracking in two dimensions. *Journal of Neurophysiology*, **81**, 1597–1602.

Ernst, M.O., & Banks, M.S. (2002). Humans integrate visual and haptic information in a statistical optimal fashion. *Nature*, **415**, 429–433.

Estes, W.K. (1976). The cognitive side of probability learning. *Psychological Review*, **83**, 37–64.

Fiser, J., & Aslin, R.N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences, USA*, **99**, 15822–15826.

Frank, M.C., Slemmer, J.A., Marcus, G.F., & Johnson, S.P. (2009). Information from multiple modalities helps 5-month-olds learn abstract rules. *Developmental Science*, **12**, 504–509.

Gaissmaier, W., & Schooler, L.J. (2008). The smart potential behind probability matching. *Cognition*, **109**, 416–422.

Gaissmaier, W., Schooler, L.J., & Rieskamp, J. (2006). Simple predictions fueled by capacity limitations: when are they successful? *Journal of Experimental Psychology: Learning, Memory and Cognition*, **32**, 966–982.

Gardner, R.A. (1957). Probability-learning with two and three choices. *American Journal of Psychology*, **70**, 174–185.

Gluck, M.A., Shohamy, D., & Myers, C. (2002). How do people solve the 'weather prediction' task? Individual variability in strategies for probabilistic category learning. *Learning & Memory*, **9**, 408–418.

Griffiths, T.L., & Tenenbaum, J.B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, **17**, 767–773.

Hudson Kam, C.L., & Newport, E.L. (2005). Regularizing unpredictable variation: the roles of adult and child learners in language formation and change. *Language Learning and Development*, **1**, 151–195.

Hudson Kam, C.L., & Newport, E.L. (2009). Getting it right by getting it wrong: when learners change languages. *Cognitive Psychology*, **59**, 30–66.

Johnson, J.S., & Newport, E.L. (1989). Critical period effects in second language learning: the influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, **21**, 60–99.

Kail, R. (1991). Processing time decreases exponentially during childhood and adolescence. *Developmental Psychology*, **27**, 259–266.

Kemp, C., Perfors, A., & Tenenbaum, J.B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, **10**, 307–321.

Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Current Opinion in Neurobiology*, **13**, 150–158.

Kidd, C., Piantadosi, S.T., & Aslin, R.N. (2012). The Goldilocks effect: human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLoS ONE*, **7**, e36399.

Kirkham, N.Z., Slemmer, J.A., & Johnson, S.P. (2002). Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition*, **83**, B35–B42.

Knowlton, B.J., Mangels, J.A., & Squire, L.R. (1996). A neostriatal habit learning system in humans. *Science*, **273**, 1399–1402.

Körding, K.P., & Wolpert, D.M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, **10**, 320–326.

Kosslyn, S.M. (1978). Measuring the visual angle of the mind's eye. *Cognitive Psychology*, **10**, 356–389.

Kovács, A.M., & Mehler, J. (2009). Cognitive gains in 7-month-old bilingual infants. *Proceedings of the National Academy of Sciences, USA*, **106**, 6556–6560.

Lunn, M., & McNeil, D. (1995). Applying Cox regression to competing risks. *Biometrics*, **51**, 524–532.

McKenzie, C.R., Lee, S.M., & Chen, K.K. (2002). When negative evidence increases confidence: change in belief after hearing two sides of a dispute. *Journal of Behavioral Decision Making*, **15**, 1–18.

McMurray, B., & Aslin, R.N. (2004). Anticipatory eye movements reveal infants' auditory and visual categories. *Infancy*, **6**, 203–229.

MacWhinney, B. (1998). Models of the emergence of language. *Annual Review of Psychology*, **49**, 199–227.

Maratsos, M. (2000). More overregularizations after all: new data and discussion on Marcus, Pinker, Ullman, Hollander, Rosen & Xu. *Journal of Child Language*, **27**, 183–212.

Marcus, G., Pinker, S., Ullman, M., Hollander, M., Rosen, T., & Xu, F. (1992). Overregularisations in language acquisition. *Monographs of the Society for Research in Child Development*, **57** (4, Serial no. 228).

Maslen, R.J.C., Theakston, A.L., Lieven, E.V.M., & Tomasello, M. (2004). A dense corpus study of past tense and plural overregularization in English. *Journal of Speech, Language, and Hearing Research*, **47**, 1319–1333.

Newport, E.L. (1990). Maturational constraints on language learning. *Cognitive Science*, **14**, 11–28.

Nisbett, R.N., Zuckier, H., & Lemley, R.E. (1981). The dilution effect: nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology*, **13**, 248–277.

Oliva, A., & Torralba, A. (2006). The role of context in object recognition. *Trends in Cognitive Sciences*, **11**, 520–527.

Pinker, S., & Ullman, M.T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, **6**, 456–463.

Plunkett, K., & Marchman, V. (1993). From rote learning to system building: acquiring verb morphology in children and connectionist nets. *Cognition*, **48**, 21–69.

Ramscar, M., Yartlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbol learning. *Cognitive Science*, **34**, 909–957.

Richardson, D.C., & Kirkham, N.Z. (2004). Multimodal events and moving locations: eye movements of adults and 6-month-olds reveal dynamic spatial indexing. *Journal of Experimental Psychology: General*, **133**, 46–62.

Rumelhart, D.E., & McClelland, J.L. (1986). On learning the past tenses of English verbs. In J.L. McClelland & D.E. Rumelhart (Eds.), *Parallel distributed processing (Vol 2): Psychological and biological models* (pp. 216–271). Cambridge, MA: MIT Press.

Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science*, **274**, 1926–1928.

Saffran, J.R., Pollack, S.D., Seibel, R.L., & Shkolnik, A. (2007). Dog is a dog is a dog: infant rule learning is not specific to language. *Cognition*, **105**, 669–680.

Shanteau, J. (1975). Averaging versus multiplying combination rules of inference judgment. *Acta Psychologica*, **39**, 83–89.

Shiffrin, R.M., & Steyvers, M. (1997). A model for recognition memory: REM: Retrieving Effectively from Memory. *Psychonomic Bulletin & Review*, **4**, 145–166.

Smith, L.B., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, **106**, 1558–1568.

Shukla, M., White, K.S., & Aslin, R.N. (2011). Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-month-old infants. *Proceedings of the National Academy of Sciences, USA*, **108**, 6038–6043.

Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J.B., & Bonatti, L.L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science*, **332**, 1054–1059.

Tenenbaum, J.B., Kemp, C., Griffiths, T.L., & Goodman, N.D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science*, **331**, 1279–1285.

Thiessen, E.D. (2010). Effects of visual information on adults' and infants' auditory statistical learning. *Cognitive Science*, **34**, 1093–1106.

Treichler, F.R. (1967). Reinforcer preference effects on probability learning by monkeys. *Journal of Comparative and Physiological Psychology*, **64**, 339–342.

Wilson, W.A., & Rollin, R.A. (1959). Two-choice behavior of rhesus monkeys in a noncontingent situation. *Journal of Experimental Psychology*, **58**, 174–180.

Wolford, G., Newman, S., Miller, M.B., & Wig, G. (2004). Searching for patterns in random sequences. *Canadian Journal of Experimental Psychology*, **58**, 221–228.

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 5012–5015.

Xu, F., & Tenenbaum, J.B. (2007). Word learning as Bayesian inference. *Psychological Review*, **114**, 245–272.

Yang, T., & Shadlen, M.N. (2007). Probabilistic reasoning by neurons. *Nature*, **447**, 1075–1080.

Yoshida, H., & Smith, L.B. (2005). Linguistic cues enhance the learning of perceptual cues. *Psychological Science*, **16**, 90–95.

Yu, A.J., & Cohen, J.D. (2009). Sequential effects: superstition or rational behavior? *Advances in Neural Information Processing Systems*, **21**, 1873–1880.

Yuki, M., Maddux, W.W., & Masuda, T. (2007). Are the windows to the soul the same in the East and West? Cultural differences in using the eyes and mouth as cues to recognize emotions in Japan and the United States. *Journal of Experimental Social Psychology*, **43**, 303–311.

Zhu, X., Rogers, T.T., & Gibson, B.R. (2009). Human Rademacher complexity. *Advances in Neural Information Processing Systems*, **21**, 2322–2330.

Zosuls, K.M., Ruble, D.N., Tamis-LeMonda, C.S., Shrout, P.E., Bornstein, M.H., & Greulich, F.K. (2009). The acquisition of gender labels in infancy: implications for gender-typed play. *Developmental Psychology*, **45**, 688–701.

Received: 26 October 2011
Accepted: 25 August 2012

Appendix

Eye-tracking details

Eye-tracking for all participants began with a calibration procedure. Adult participants were calibrated using nine points, one at each point of a three by three grid. To expediate calibration, infant participants were calibrated using five points: the four corners and the center.

The Tobii eye tracker recorded participants' distance from the screen and the location of both their left and right eyes at 50 Hz. Each participants' sequence of gaze points was derived from their recorded gaze samples. If the Tobii x and y coordinates for both eyes were on the screen, gaze was estimated to be at their midpoint. If the coordinates of only one eye were reliably recorded, those coordinates were estimated to be the point of gaze. Otherwise, the sample was marked as invalid. Distance was treated similarly. In order to correct for blinking or other sporadic tracking failures, we interpolated over short intervals of invalid samples. Up to three successive invalid samples between two valid samples were interpolated in equal steps. Larger blocks of invalid samples were not interpolated.

Finally, these time/x/y/distance–tuples were used to estimate a series of fixations for each participant. Successive samples were considered part of the same fixation if they were within 1° of visual angle of each other (using the arctangent computation described by Kosslyn (1978) to convert from on-screen distances to visual angles) and their summed duration was greater than 100 milliseconds. Table A1 shows the proportion of time during the 1.2s prediction window that participants' fixation locations were on-screen.

Table A1 *Proportion of valid eye-tracking samples in the prediction window for participants in each experiment. Boxes show mean (std err.)*

Participants	Exp 1: Adults	Exp 1: Infants	Exp 2: Infants
Prop. valid samples	.76 (.03)	.69 (.04)	.69 (.03)